Statistica Sinica **6**(1996), 331-345

EFFECT OF THRESHOLD RULES ON PERFORMANCE OF WAVELET-BASED CURVE ESTIMATORS

Peter Hall and Prakash Patil

Australian National University

Abstract. Wavelet-based curve estimators have received considerable recent attention, particularly in terms of their ability to adapt to irregularities in a curve. Nevertheless, the threshold rules on which wavelet estimators are based are not well understood, and indeed some contemporary workers employ rules that are suboptimal by an order of magnitude. In this paper we give necessary conditions and sufficient conditions on the form of the threshold for the resulting curve estimator to achieve optimal convergence rates in the case of smooth and piecewise-smooth functions. We discuss practical threshold rules that achieve optimal rates and study spatially adaptive rules that permit a degree of local smoothing. We address the important statistical problem of which tuning parameters in threshold rules produce genuine statistical smoothing in the sense of allowing adjustment of variance against bias in a first-order sense. Some tuning parameters affect only bias while others influence neither bias nor variance to first order.

Key words and phrases: Bias, convergence rate, mean squared error, smoothing parameter, threshold, variance, wavelet.

1. Introduction

A wavelet estimator of a function f is typically based on replacing wavelet coefficients by their respective empirical versions in a formula for the "wavelet transform" of f. Now, the wavelet transform is an infinite double series over both frequency (i.e. resolution) and location, and direct substitution of coefficient estimators for true coefficients produces a double series which does not converge. In practice this difficulty is overcome by truncating the sum over resolution and including in the double series only those empirical wavelet coefficients which exceed a certain threshold. The truncation operation is relatively innocuous and is robust against adjustments. In practice it serves only to ensure that coefficient estimates that are included in the curve estimator are based on sufficiently many data values for their variability not to be excessive. However, selection of threshold can be crucial to performance. The aim of this paper is to address the issue of threshold choice in the context of smooth and piecewise-smooth function estimation.

We show that a wavelet-based curve estimator achieves optimal convergence rates if and only if the threshold which it employs is bounded by a constant multiple of $n^{-1/2}$, where n denotes sample size, up to a certain critical resolution level k_0 (depending on both n and the order of wavelet); and, beyond that point, increases at least as fast as const. $n^{-1/2}(k-k_0)^{1/2}$ among higher resolution levels k. The critical resolution level k_0 is given by an integer close to $(2r+1)^{-1}\log_2 n$, where r denotes the order of the wavelet and $\log_2$ indicates logarithm to base 2. Other threshold rules, such as the "pure threshold" approach with threshold set equal to const. $n^{-1/2}(\log n)^{1/2}$ at all resolution levels, can have creditable performance but do not quite achieve the level of accuracy of estimators with appropriately varying thresholds. The "pure threshold" estimator is intrinsically oversmoothed — it excludes too many low-resolution wavelet coefficients with the result that the bias contribution to its mean integrated squared error dominates the variance contribution.

We point out that if the threshold is selected in a way which is adaptive relative to both spatial location and resolution then variance may be balanced against bias in a spatial sense. In this case the mean integrated squared error formula is strikingly similar to its counterpart in the context of variable bandwidth kernel estimation. For example, an r'th order kernel estimator $\tilde f$ of an r-times differentiable probability density f, using the variable bandwidth $h(x) = h_0\,H(x)$ where $h_0 \simeq n^{-1/(2r+1)}$ depends only on n and $H(x)$ depends only on x, has mean integrated squared error given by

$$\int E(\tilde f - f)^2 \sim B_1(nh_0)^{-1}\int fH^{-1} + B_2\,h_0^{2r}\int f^{(r)^2}\,H^{2r}\,. \tag{1.1}$$

(The constants B_1 and B_2 depend only on the kernel function. See Rosenblatt (1971) for discussion of related results.) An identical formula holds for appropriately thresholded r'th order wavelet estimators; there, $B_1 \equiv 1$, B_2 depends only on the wavelet function, $h_0 = 2^{-k_0}$, and $\log_2 H$ must be integer-valued. The latter restriction results from the dyadic definition of scale at successive resolution levels in the wavelet transform. An important aspect of the wavelet version of (1.1) is that it continues to hold when f is only piecewise smooth. By way of contrast, kernel estimators can be seriously affected by discontinuities in f, and formula (1.1) fails there unless $r = 1$.

These considerations raise the issue of which tunable parameters in the wavelet estimator provide genuine statistical smoothing, to first order, and which parameters adjust only bias (without appreciably affecting variance) or simply ensure that the estimator is well-defined by guaranteeing convergence of the double series used to construct it. The truncation parameter is of the latter type — the manner in which it is typically used does not provide any trade-off of

bias against variance, and so does not produce statistical smoothing. Neither, in the case of the pure threshold estimator noted earlier, does the constant in the threshold formula "const. $(n^{-1} \log n)^{1/2}$" provide any statistical smoothing to first order. For all permitted choices of that constant, adjustments to its value affect only bias to first order. The pure threshold estimator does not have a tunable smoothing parameter. However, more flexible threshold rules do offer a real potential for statistical smoothing. Such smoothing may be available globally — for example, the critical resolution level discussed in the second paragraph of this section provides global smoothing — or locally as in the case of a spatially varying threshold.

We should stress that these conclusions are founded on a specific model for the true function f, which (when it is estimated using an r'th order wavelet) is taken to have r bounded derivatives in a piecewise sense. Here the spatial adaptability of wavelet methods is evident from their ability to cope with discontinuities in derivatives of lower order than r, and also (in the case of spatially varying thresholds) their potential for providing local smoothing in this context. In this respect more traditional methods, such as those based on kernels, do not perform so well, and in fact produce estimators whose convergence rates are an order of magnitude slower than those of wavelet estimators when f has discontinuities. Generally, the pure threshold method responds relatively well to high-order episodes in the curve, such as fluctuations whose frequency is of larger size than $(\log n)^{1/2}$. This issue has been taken up by Hall and Patil (1995a) and Fan, Hall, Patil and Martin (1993). Nevertheless, a critical point made by the present paper is that pure thresholding does not adapt at all well to the lower frequency changes with which statistical scientists are often concerned.

Wavelet methods were introduced to statistics by Donoho (1992), Donoho and Johnstone (1992a-b, 1994) and Kerkyacharian and Picard (1992, 1992a-c). The approach adopted in the present paper differs from that taken by these authors in that we focus attention on a fixed target function f rather than describe performance uniformly over large classes of f's. The purpose of our restriction is to identify more clearly the ways in which choice of threshold affects the performance of a wavelet-based curve estimator. There is no difficulty in generalizing our results so that they apply uniformly over classes of functions with r bounded derivatives in a piecewise sense. However, in most cases our claims about the performance of different threshold rules, and the extent to which they provide genuine statistical smoothing, are not valid uniformly over the very large Sobolev spaces considered by Donoho, Johnstone, Kerkyacharian and Picard.

The results produced in this paper are, of course, theoretical, and the arguments that derive them are heavily mathematical. To avoid clouding the central issues by technical matters we give all derivations in outline only. Concise details

follow lines pursued by Hall and Patil (1995b), to which the reader interested in the technical side is referred. As in that paper we address only the case of nonparametric density estimation, in the knowledge that other applications, such as nonparametric regression, do not differ in qualitative terms. Thus, our results about the efficacy of different threshold rules are available quite generally.

Organization of the rest of the paper is as follows. Our main results, as indicated in the second paragraph above, are described in Section 2. In Section 3 we give three specific examples of threshold rules. The first two achieve the optimal mean square convergence rate, although the third — pure thresholding — does not. Still in Section 3 we propose refinements of those threshold rules which achieve optimal mean square convergence rates. We also discuss spatially adaptive threshold rules. The issue of which tunable parameters provide genuine statistical smoothing is discussed in Section 4. Extensions of our results and conclusions to piecewise-smooth densities are indicated in Section 5.

2. Main Results

2.1. Summary, and introduction to wavelet transforms

Section 2.2 introduces approximations to wavelet coefficients and their estimators, and discusses some of the implications of those formulae. The formulae are applied in Section 2.3 to derive necessary and sufficient conditions for r'th order wavelet-based density estimators to achieve the optimal convergence rate, $O(n^{-2r/(2r+1)})$ among densities in a class of r-times differentiable functions,

$$\mathcal{F}_r(B) = \Big\{ f \ge 0 : \int f = 1\,,\ \sup |f^{(j)}| \le B\,, \int |f^{(j)}| \le B \text{ for } 0 \le j \le r\,,\ f^{(r)} \text{ is uniformly continuous on } (-\infty,\infty)\,, \text{ and } f^{(r)} \text{ is monotone on } (-\infty,-B) \text{ and on } (B,\infty) \Big\}\,,$$

where $B \ge 1$. The monotonicity assumption here is employed to enable infinite series to be approximated by integrals, in calculation of bias terms. It seems so mild that we have not attempted to relax it.

Next we describe the wavelet transform and its empirical version for nonparametric density estimation. Let ψ and ϕ denote respectively mother and father wavelet functions of r'th order, enjoying the properties $\int \phi^2 = \int \psi^2 = 1$, $\int x^i\,\psi(x)\,dx = 0$ for $1 \le i \le r-1$, and $= r!\kappa$ (say) $\ne 0$ when $i = r$. Furthermore, for arbitrary $p > 0$, and defining $p_k = p\,2^k$ for $k \ge 0$, the functions $\phi_l(x) = p^{1/2}\,\phi(px - l)$ and $\psi_{kl}(x) = p_k^{1/2}\,\psi(p_k x - l)$ form an orthonormal basis for the class of square-integrable functions f. The orthogonality relations may be expressed by $\int \phi_{l_1}\,\phi_{l_2} = \delta_{l_1 l_2}$, $\int \psi_{k_1 l_1}\,\psi_{k_2 l_2} = \delta_{k_1 k_2}\,\delta_{l_1 l_2}$, $\int \phi_{l_1}\,\psi_{k l_2} = 0$, where

δ_{ij} denotes the Kronecker delta. The wavelet transform of a square-integrable function f is given by

$$f = \sum_l b_l \, \phi_l + \sum_{k=0}^{\infty} \sum_l b_{kl} \, \psi_{kl} \,, \tag{2.1}$$

where $b_l = \int f \, \phi_l$ and $b_{kl} = \int f \, \psi_{kl}$ are wavelet coefficients.

In addition to the standard properties of wavelets listed above, we suppose that both ϕ and ψ are bounded and compactly supported.

If f is a probability density and $X_1, \ldots, X_n$ is a random sample from the associated probability distribution, then

$$\hat{b}_l = n^{-1} \sum_{i=1}^{n} \phi_l(X_i) \quad \text{and} \quad \hat{b}_{kl} = n^{-1} \sum_{i=1}^{n} \psi_{kl}(X_i)$$

are unbiased estimators of b_l and b_{kl}, respectively. An empirical version of (2.1) is given by

$$\hat{f} = \sum_l \hat{b}_l \, \phi_l + \sum_{k=0}^{q-1} \sum_l \hat{b}_{kl} \, I(|\hat{b}_{kl}| > t_{kl}) \, \psi_{kl} \,, \tag{2.2}$$

where $q \geq 1$ is a truncation parameter, t_{kl} is a threshold and I is the indicator function. If we were to remove the truncation and threshold operations — that is, take $q = \infty$ and $t_{kl} = 0$ — then the double series at (2.2) would not converge.

These representations of f and $\hat{f}$ are not unique. In particular, if k^* is any positive integer then the function f may be expressed as

$$f = \sum_l b^*_l \, \phi^*_l + \sum_{k=0}^{\infty} \sum_l b^*_{kl} \, \psi^*_{kl} \,,$$

where $\phi^*_l(x) = (p\,2^{k^*})^{1/2} \, \phi(p\,2^{k^*} x - l)$, $\psi^*_{kl} = \psi_{k^*+k,l}$ $b^*_l = \int f \, \phi^*_l$ and $b^*_{kl} = b_{k^*+k,l}$. Furthermore, if $t_{kl} = 0$ whenever $0 \leq k \leq k^* - 1$ then the estimator $\hat{f}$ defined at (2.2) is identical to

$$\hat{f} = \sum_l \hat{b}^*_l \, \phi^*_l + \sum_{k=0}^{q-k^*-1} \sum_l \hat{b}^*_{kl} \, I(|\hat{b}^*_{kl}| > t_{k^*+k,l}) \, \psi^*_{kl} \,, \tag{2.3}$$

where $\hat{b}^*_l = n^{-1} \sum_{i=1}^n \phi^*_l(X_i)$ and $\hat{b}^*_{kl} = n^{-1} \sum_{i=1}^n \psi^*_{kl}(X_i) = \hat{b}_{k^*+k,l}$. In view of these different expressions for the same estimator it is a little awkward to discuss thresholds unambiguously, particularly since there exist important classes of estimators that have all low-resolution thresholds equal to zero. We remove this difficulty by taking $p_k \equiv \xi \, 2^k$, where $\xi \in [1, 2)$. Much of the treatment of wavelet methods extant in the literature is for the case $\xi = 1$, and to simplify

our discussion we adopt this convention in Sections 2.3 and 3.1. However, such an approach allows smoothing via the resolution level to be implemented only in discrete, dyadic steps. It does not avail itself of the continuum of smoothing that may be enjoyed by adjusting ξ as well as k, as we show in Section 3.2.

For the sake of simplicity we adhere to convention and define k to be the resolution level associated with the contribution $\hat{b}_{kl}\, I(|\hat{b}_{kl}| > t_{kl})\, \psi_{kl}$ to the estimator $\hat{f}$. However, this is not entirely satisfactory since adjoining the variable ξ, as discussed above, does slightly alter the effective resolution. Hall and Patil (1995a,b) denoted resolution by p_k, rather than k, to avoid this problem.

We shall assume throughout that the truncation point, q, is chosen so that $2^q = O(n^{1-\epsilon})$ for some $\epsilon > 0$.

2.2. Approximations to b_{kl}, $\hat{b}_{kl}$ and mean integrated squared error

Observe that, by Parseval's identity, the mean integrated squared error of the estimator defined at (2.2) is given by

$$\text{MISE} = \int E(\hat{f} - f)^2 = \sum_l a_l + \sum_{k=0}^{q-1} \sum_l (a_{1kl} + a_{2kl}) + \sum_{k=q}^{\infty} \sum_l b_{kl}^2\,, \tag{2.4}$$

where $a_l = E(\hat{b}_l - b_l)^2$, and

$$a_{1kl} = E\{(\hat{b}_{kl} - b_{kl})^2\, I(|\hat{b}_{kl}| > t_{kl})\}\,, \quad a_{2kl} = b_{kl}^2\, P(|\hat{b}_{kl}| \le t_{kl})\,.$$

It is not difficult to see, for plausible choices of the threshold t_{kl}, that a_{2kl} dominates a_{1kl} when k is very large, and a_{1kl} dominates a_{2kl} when k is small. The approximations that we develop below are for the intermediate case, where a_{1kl} and a_{2kl} are of similar sizes, or at least have a ratio that is not excessively large or small. Values of k (and l) which produce this intermediate behaviour are those which determine the performance of thresholding rules.

With the exception of Section 3.3, and small portions of Sections 4 and 5, we suppose below that the threshold t_{kl} depends only on k. This is typically the case in practice, and by making that assumption we may simplify both our notation and our discussion. To express the assumption we write the threshold as t_k.

Let $\theta = \theta(k, l, x)$ denote a quantity lying between 0 and 1, and put $\beta = \kappa\, f^{(r)}$, a bounded function. Then for $f \in \mathcal{F}_r(B)$,

$$\begin{aligned} b_{kl} = E(\hat{b}_{kl}) &= p_k^{-1/2} \int \psi(x)\, f\{(x+l)/p_k\}\, dx \\ &= p_k^{-1/2} \int \psi(x) \Big[\sum_{j=0}^{r-1} (j!)^{-1} (x/p_k)^j\, f^{(j)}(l/p_k) \\ &\qquad + (r!)^{-1} (x/p_k)^r\, f^{(r)}\{(\theta x + l)/p_k\} \Big]\, dx \end{aligned}$$

$$= p_k^{-\{r+(1/2)\}}\,(r!)^{-1}\int x^r\,\psi(x)\,f^{(r)}\{(\theta x+l)/p_k\}\,dx$$
$$\simeq p_k^{-\{r+(1/2)\}}\,\beta(l/p_k)\,.$$

Also,

$$n\ \mathrm{Var}(\hat b_{kl}) = E\{\psi_{kl}(X)^2\} - b_{kl}^2 \simeq E\{\psi_{kl}(X)^2\} = \int \psi(x)^2\,f\{(x+l)/p_k\}\,dx$$
$$\simeq f(l/p_k)\,,$$

and, provided n/p_k is large, $\hat b_{kl}$ is approximately Normally distributed. Therefore, writing N for a Normal $N(0,1)$ random variable,

$$a_{1kl} \simeq n^{-1}\,f(l/p_k)\,E[N^2\,I\{|n^{-1/2}\,f(l/p_k)^{1/2}N + p_k^{-\{r+(1/2)\}}\,\beta(l/p_k)| > t_k\}]\,,$$
$$a_{2kl} \simeq p_k^{-(2r+1)}\,\beta(l/p_k)^2\,P\{|n^{-1/2}\,f(l/p_k)^{1/2}N + p_k^{-\{r+(1/2)\}}\,\beta(l/p_k)| \le t_k\}\,.$$

Similarly,

$$\sum_{k=q}^{\infty}\sum_l b_{kl}^2 \sim \Big(\int\beta^2\Big)\sum_{k=q}^{\infty}p_k^{-2r} = (1-2^{-2r})^{-1}\Big(\int\beta^2\Big)\,p_q^{-2r}\,.$$

Also, for compactly supported ϕ, $\sum_l a_l = O(n^{-1})$. Arguing thus, writing $t_k = n^{-1/2}\,\lambda_k$ for a positive constant λ_k (possibly depending on n as well as k), and noting (2.4) and the fact that the MISE is of larger order than n^{-1}, we deduce that

$$\begin{aligned}\mathrm{MISE} \sim \sum_{k=0}^{q-1}\int\Big(&n^{-1}\,p_k\,f(x)\,E[N^2\,I\{|n^{-1/2}\,f(x)^{1/2}\,N\\ &\qquad + p_k^{-\{r+(1/2)\}}\,\beta(x)| > n^{-1/2}\,\lambda_k\}]\\ &+ p_k^{-2r}\,\beta(x)^2\,P\{|n^{-1/2}\,f(x)^{1/2}\,N\\ &\qquad + p_k^{-\{r+(1/2)\}}\,\beta(x)| \le n^{-1/2}\,\lambda_k\}\Big)\,dx\\ &+(1-2^{-2r})^{-1}\Big(\int\beta^2\Big)\,p_q^{-2r}\,.\end{aligned}\tag{2.5}$$

2.3. Choice of threshold and truncation when $p_k \equiv 2^k$

Let k_0 denote the integer part of $(2r+1)^{-1}\log_2 n = (2r+1)^{-1}(\log 2)^{-1}(\log n)$, this quantity being chosen since it ensures that $p_{k_0}^{2r+1}/n$ is bounded away from 0 and ∞ as $n\to\infty$. Let $C_1, C_2, \ldots$ be positive constants, depending on f but not on k or n. We claim that if $f\in\mathcal{F}_r(B)$ then in order for $\mathrm{MISE} = O(n^{-2r/(2r+1)})$ it is

(a) necessary that there exist $C_1,\ldots,C_5$ such that for all n: $q \ge k_0 - C_1$, $\#\{k : 0\le k\le k_0-1$, and $\lambda_k > C_2\} \le C_3$, and $\#\{k : k_0\le k\le q-1 : \lambda_k \le C_4(k-k_0)^{1/2}\} \le C_5$; and

(b) sufficient that there exist $C_1, \ldots, C_5$ with C_4 chosen sufficiently large, such that for all n : $q \geq k_0 - C_1$, $\#\{k : 0 \leq k \leq k_0 - 1, \lambda_k > C_2\} \leq C_3$, and $\#\{k : k_0 \leq k \leq q-1, \lambda_k \leq C_4(k-k_0)^{1/2}\} \leq C_5$.

To establish these claims first observe that if $p_k \equiv 2^k$ then the very last term on the right-hand side of (2.5) equals $O(n^{-2r/(2r+1)})$ if and only if, for some $C_6 > 0$, $2^q \geq C_6\, n^{1/(2r+1)}$; or equivalently, if and only if $q \geq k_0 - C_7$.

Next, note that the contribution to the right-hand side of (2.5) from that part of the series corresponding to the sum over $k_0 \leq k \leq q-1$, dominates

$$C_8\, n^{-2r/(2r+1)} \sum_{k=0}^{q-k_0-1} 2^k\, E\{N^2\, I(|N| > C_9\, \lambda_{k_0+k})\}$$
$$\geq C_{10}\, n^{-2r/(2r+1)} \sum_{k=0}^{q-k_0-1} 2^k\, \exp\,(\,-C_{11}\, \lambda_{k_0+k}^2)$$

and is dominated by

$$C_{12}\, n^{-2r/(2r+1)} \sum_{k=0}^{q-k_0-1} 2^k\, \exp\,[\,-C_{13}\{\max(\lambda_{k_0+k} - C_{14}, 0)\}^2]\,.$$

Similarly, the contribution from the sum over $0 \leq k \leq k_0 - 1$ dominates

$$C_{15}\, n^{-2r/(2r+1)} \sum_{k=0}^{k_0-1} \{I(2^{k\{r+(1/2)\}} \leq C_{16}\, \lambda_k) - P(|N| > C_{17}\, 2^{k\{r+(1/2)\}})\}$$

and is dominated by

$$C_{18}\, n^{-2r/(2r+1)} \sum_{k=0}^{k_0-1} \{I(2^{k\{r+(1/2)\}} \leq C_{19}\, \lambda_k) + P(|N| > C_{20}\, 2^{k\{r+(1/2)\}})\}\,.$$

The claims two paragraphs above follow from these bounds. Indeed, the bounds may be established uniformly in densities $f \in \mathcal{F}_r(B)$, even with the assumption that $f^{(r)}$ is continuous removed from the definition of $\mathcal{F}_r(B)$; and so conditions (a) and (b) above are respectively necessary and sufficient for achieving the optimal convergence rate *uniformly* in $f \in \mathcal{F}_r(B)$.

3. Threshold Rules

3.1. Summary and examples

In the present section we give three specific examples of threshold rules and discuss them in the light of results in Section 2.3. Section 3.2 presents refinements of the first two rules, which produce optimal mean square convergence rates, and briefly addresses the issue of soft thresholding. Spatially adaptive thresholds are described in Section 3.3, and their properties outlined.

Let $p_k \equiv 2^k$ and write $C_1, C_2, \ldots$ for positive constants. In view of the results derived in Section 2.3, the following two threshold rules achieve the optimal mean square convergence rate $n^{-2r/(2r+1)}$ uniformly over densities in $\mathcal{F}_r(B)$. Assume that $k_0 - C_1 \le q \le C_2(\log n)(\log 2)^{-1}$ for arbitrary $C_1 > 0$ and $C_2 \in ((2r+1)^{-1}, 1)$. Take $\lambda_k = 0$ for $0 \le k \le k_1$, and either $\lambda_k = C_3(k-k_1)^{1/2}$ for $k > k_1$, or $\lambda_k = C_3(\log n)^{1/2}$ for $k > k_1$, where $|k_1 - k_0| \le C_4$, $C_4 > 0$ is arbitrary, and C_3 is sufficiently large.

However, a third threshold rule, given by $\lambda_k \equiv C_5(\log n)^{1/2}$ for all k, does not produce the mean square convergence rate $n^{-2r/(2r+1)}$. No matter what the value of C_5 the rate is no better than $(n^{-1}\log n)^{2r/(2r+1)}$, as may be deduced from (2.5). The reason for the logarithmic factor is that the density estimator with this fixed threshold is oversmoothed in the sense that the variance contribution to MISE is asymptotically negligible relative to the squared bias contribution. To illustrate this point we treat the case of large C_5 in a little detail. With $C_5 > (2 \sup f)^{1/2}$ and $q \sim C_2(\log n)(\log 2)^{-1}$ where $C_2 \in ((2r+1)^{-1}, 1)$, it may be shown that

$$\sum_{k=0}^{q-1} n^{-1} p_k \int f(x)\, E[N^2\, I\{|n^{-1/2} f(x)^{1/2} N + p_k^{-\{r+(1/2)\}} \beta(x)| > n^{-1/2}\lambda_k\}]\, dx$$
$$= o\{(n^{-1}\log n)^{2r/(2r+1)}\}, \tag{3.1}$$

$$\sum_{k=0}^{q-1} p_k^{-2r} \int \beta(x)^2\, P\{|n^{-1/2} f(x)^{1/2} N + p_k^{-\{r+(1/2)\}} \beta(x)| \le n^{-1/2}\lambda_k\}\, dx$$
$$\asymp (n^{-1}\log n)^{2r/(2r+1)}, \tag{3.2}$$

and $p_q^{-2r} = o\{(n^{-1}\log n)^{2r/(2r+1)}\}$, in which the notation "$a_n \asymp b_n$" means "a_n/b_n and b_n/a_n are both bounded sequences". (It is possible to refine (2.7) by replacing the relation "$\asymp (n^{-1}\log n)^{2r/(2r+1)}$" by "$= c_n(n^{-1}\log n)^{2r/(2r+1)}$", but the sequence $\{c_n\}$ here does not converge, although its lim inf is strictly positive and its lim sup finite.) Therefore, in view of (2.5),

$$\text{MISE} \asymp (n^{-1}\log n)^{2r/(2r+1)}.$$

Now, the integral of squared bias of $\hat{f}$ is asymptotic to the left-hand side of (3.2) — all but a negligibly small part of the integral of variance is expressed by the left-hand side of (3.1). Therefore, the threshold choice $\lambda_k \equiv C_5(\log n)^{1/2}$ renders the estimator oversmooth, with squared bias dominating variance.

3.2. Refinements in the case of "optimal" threshold rules

We begin by addressing the first two threshold rules suggested in Section 3.1, and refine them by taking $p_k = \xi\, 2^k$, where $\xi \in [1, 2)$ ranges over a continuum of

values. The respective thresholds are given by $t_k = n^{-1/2}\,\lambda_{1,k}$ and $t_k = n^{-1/2}\,\lambda_{2,k}$, where

$$\lambda_{1,k} = \begin{cases} 0, & \text{if} \quad 0 \le k \le k_1, \\ C(k-k_1)^{1/2}, & \text{if} \quad k > k_1, \end{cases} \tag{3.3a}$$

and

$$\lambda_{2,k} = \begin{cases} 0, & \text{if} \quad 0 \le k \le k_1, \\ C(\log\, n)^{1/2}, & \text{if} \quad k > k_1, \end{cases} \tag{3.3b}$$

and $C > 0$. It is easily checked that the results derived in Section 2.3, concerning necessary and sufficient conditions for optimal performance of the thresholded estimator, continue to apply in this slightly more general setting.

For simplicity, let q equal the integer part of $C'(\log\, n)(\log\, 2)^{-1}$ where $C' \in ((2r+1)^{-1}, 1)$. Define $p = p_{k_1} = \xi\, 2^{k_1}$, $\phi_l(x) = p^{1/2}\phi(px - l)$, $b_l = \int f\,\phi_l$, $\hat{b}_l = n^{-1}\,\sum_{j=1}^n \phi_l(X_j)$. In this notation the formulae at (2.1) and (2.2) may be equivalently expressed as

$$\begin{aligned} f &= \sum_l b_l\,\phi_l + \sum_{k=0}^{\infty} \sum_l b_{kl}\,\psi_{kl}\,, \\ \hat{f}_j &= \sum_l \hat{b}_l\,\phi_l + \sum_{k=0}^{q-k_1-1} \sum_l \hat{b}_{kl}\,I(|\hat{b}_{kl}| > n^{-1/2}\,\lambda_{j,k_1+k+1})\,\psi_{kl}\,. \end{aligned} \tag{3.4}$$

Result (2.5) may be employed to show that, again with $p = \xi\, 2^{k_1}$,

$$\int E(\hat{f}_j - f)^2 \sim n^{-1}\,p + p^{-2r}\,a_j(f), \tag{3.5}$$

where

$$\begin{aligned} a_1(f) =& \sum_{k=1}^{\infty} \Big(\rho\, 2^k \int f(x)\, E[N^2\, I\{|f(x)^{1/2}\,N + \rho^{-1/2}\,2^{-(2r+1)k/2}\,\beta(x)| > C\,k^{1/2}\}]\,dx \\ &+ 2^{-2rk} \int \beta(x)^2 P\{|f(x)^{1/2}N + \rho^{-1/2}2^{-(2r+1)k/2}\beta(x)| \le Ck^{1/2}\}dx \Big), \\ a_2(f) =& (1 - 2^{-2r})^{-1} \int \beta^2\,, \end{aligned}$$

$\rho = \rho(n) = n^{-1}\,p^{2r+1}$ (which is bounded away from 0 and ∞ if $|k_1 - k_0| \le C_1$), and it is assumed that C is sufficiently large. (For both threshold rules, $C > \{2(\log\, 2)\,\sup\, f\}^{1/2}$ is adequate.)

It is intuitively clear, and also follows from the definitions of $a_1(f)$ and $a_2(f)$, that the ratio of the asymptotic mean integrated squared errors of $\hat{f}_1$ and $\hat{f}_2$ may be rendered arbitrarily close to 1 by choosing C sufficiently large. Of course, the

asymptotic mean integrated squared error of $\hat{f}_2$ does not depend on C. Therefore, in sheer asymptotic terms the estimator $\hat{f}_2$ is not inferior to $\hat{f}_1$, and generally surpasses it for appropriate choice of C and p.

Formula (3.5) is an analogue of the more familiar mean integrated squared error formula for an r'th order kernel density estimator,

$$\tilde{f}(x) = (nh)^{-1} \sum_{j=1}^{n} K\{(x - X_j)/h\}\,,$$

where K satisfies $\int x^i K(x)\,dx = 1$ for $i = 0$, and $= 0$ for $1 \le i \le r-1$. There,

$$\int E(\tilde{f} - f)^2 \sim (nh)^{-1} \Big(\int K^2 \Big) + h^{2r} \Big\{ \int (f^{(r)})^2 \Big\} \Big\{ \int x^r\, K(x)\,dx \Big\}^2\,, \qquad (3.6)$$

with which (3.5) compares if we consider p^{-1} to be the analogue of bandwidth, h.

The threshold rules employed to construct $\hat{f}_1$ and $\hat{f}_2$ are called "hard thresholds", in that they either include or exclude all of $\hat{b}_{kl}$ — there are no half measures. A more general rule has the form

$$\hat{f} = \sum_l \hat{b}_l\, \phi_l + \sum_{k=k_1+1}^{q-1} \sum_l \hat{b}_{kl}\, w(n^{1/2}\, |\hat{b}_{kl}|/\lambda_k)\, \psi_{kl}$$

(compare (3.4)), where the function w satisfies $w(u) = 0$ for $0 < u < c_1$, $w(u) \in [0,1]$ for $c_1 \le u \le c_2$, and $w(u) = 1$ if $u > c_2$, with $0 < c_1 \le c_2 < \infty$ being constants. "Hard thresholding" has $c_1 = c_2 = 1$; "soft thresholding" has $c_1 < c_2$ and w continuous. Analogues of the mean integrated squared error formula above are readily established for soft threshold rules. There are no changes to the conclusions that we have drawn. Generally, w influences mean squared error formulae only through terms of second order, and so we do not deal here with the issue of optimal selection of w.

3.3. Spatially adaptive thresholds

Throughout the discussion above we have taken the threshold $t = t_k$ to be adaptive only with respect to resolution, expressed by k. The threshold has not been adapted to location, i.e. spatial position, expressed by l. However, we might allow t to depend on both k and l. For example, consider a spatially adaptive version of the second of the two threshold rules discussed in Sections 3.1 and 3.2, where λ_k was either 0 or $C(\log n)^{1/2}$. Re-define

$$\lambda_k = \lambda_k(l) = \begin{cases} 0, & \text{if} \quad 0 \le k \le k_1(l), \\ C(\log n)^{1/2}, & \text{if} \quad k > k_1(l)\,, \end{cases} \qquad (3.7)$$

where $k_1(l) = k_0 + g(l/p_0)$, g is an integer-valued function, $p_0 = \xi\, 2^{k_0}$, $\xi \in [1,2)$, and $C > \{2(\log 2)\ \sup f\}^{1/2}$. The corresponding wavelet density estimator is

$$\hat f = \sum_l \hat b_l\, \phi_l + \sum_{k=0}^{q-1} \sum_l \hat b_{kl}\, I\{|\hat b_{kl}| > n^{-1/2}\, \lambda_k(l)\}\, \psi_{kl}\,, \tag{3.8}$$

and in view of (2.5) its mean integrated squared error satisfies

$$\int E(\hat f - f)^2 \sim n^{-1}\, p_0\, A + p_0^{-2r}\, B\,, \tag{3.9}$$

where

$$A = \int f\, 2^g\,, \qquad B = (1 - 2^{-2r})^{-1} \int \beta^2\, 2^{-2rg}\,. \tag{3.10}$$

In practice one might construct pilot estimators of f and $f^{(r)}$, employ those to calculate estimates of A and B as functionals of g, and thereby compute an empirical approximation to the function g which minimizes the right-hand side of (3.9).

There is of course a version of this estimator defined by analogy with $\hat f_1$. (The latter was introduced in Section 3.2.) To define it, replace the formula at (3.7) by

$$\lambda_k = \lambda_k(l) = \begin{cases} 0, & \text{if} \quad 0 \le k \le k_1(l), \\ C\{k - k_1(l)\}^{1/2}, & \text{if} \quad k > k_1(l)\ , \end{cases} \tag{3.11}$$

where again $k_1(l) = k_0 + g(l/p_0)$, and with this change, let $\hat f$ be given by (3.8). Formula (3.9) holds for the new estimator, with the same expression for A as before but a new, more complex definition of B.

More generally still, in formula (3.11) one could take C to be a function of location, of the form $C(l/p_0)$ but constrained to exceed $\{2(\log 2)\ \sup f\}^{1/2}$. In practice one might wish to vary just one of k_1 and C, but not both, with location.

In the context of spatially adaptive rules, it is of technical interest to note that if λ_k is taken to be a function of x then ideally one should choose $\lambda_k(x) = 0$ or ∞ according as $\beta(x)^2 f(x)^{-1} >$ or $< n^{-1} p_k^{2r+1}$. This follows from (2.5) and the fact that the quantity

$$E\{N^2 I(|N + \mu| > \lambda)\} + \mu^2 P(|N + \mu| \le \lambda)$$

is minimized by taking $\lambda = 0$ or ∞ according as $|\mu| >$ or < 1. However, in practice it is not feasible to implement a threshold rule with this degree of precision. Furthermore, much less elaborate rules (e.g. those not depending on spatial location) enjoy identical convergence rates.

4. Smoothing Parameters

Here we address the issue of which tuning parameters produce genuine statistical smoothing, and which have other roles. We define a smoothing parameter to be a variable which may be adjusted to effect *first-order* changes in both variance and squared bias contributions to mean integrated squared error, such that one of these quantities increases while the other decreases. All the tuning parameters involved in the definitions of our wavelet estimators have some impact on both bias and variance, but the effect is often only of second order. For the sake of brevity we focus on "hard" thresholding rules.

The simplest case is perhaps that of the estimator $\hat{f}_2$, defined at (3.4). There the variance and (squared) bias contributions to asymptotic mean integrated squared error are given respectively by the first and second terms on the right-hand side of (3.5). They depend only on p. Increasing p increases variance but decreases bias. Thus, p is the only smoothing parameter. In particular, the constant C appearing in the definition of the threshold $\lambda_{2,k}$ is not a smoothing parameter. Choosing it within the permitted range does not lead to any first-order changes to either the variance or bias contributions to mean integrated squared error.

The case of the estimator $\hat{f}_1$ is more complex. There, the second term on the right-hand side of (3.5) involves contributions from both variance and bias — compare the case of $\hat{f}_2$, where it derived solely from bias. The threshold constant C enters through the value of $a_1(f)$, and there C does effect a trade-off between bias and variance. Thus, both p and C are smoothing parameters in this example.

We turn next to the "pure threshold estimator" discussed in the latter part of Section 3.1, which uses the threshold $\lambda_k \equiv C_5(\log\, n)^{1/2}$ for all k. As noted earlier, asymptotic mean integrated squared error in this case is, for $C_5 > C_5^0$ say, dominated entirely by squared bias — the estimator is significantly oversmoothed. Therefore, while adjustment of C_5 within the permitted range influences the level of bias, it does not effect a trade-off between variance and bias contributions, at least to first order. The constant C_5 is therefore not a smoothing parameter.

Next we examine the first of the two spatially adaptive threshold rules discussed in Section 3.3; see (3.7) and (3.8). There, as in the case of $\hat{f}_2$, influence of the constant C in the definition of the threshold vanishes entirely, in first-order terms. The only smoothing parameters are p_0 and the integer-valued function g, which effects a certain amount of local adaptive smoothing. The variance contribution to asymptotic mean integrated squared error is given by the first term on

the right-hand side of (3.9), and the bias contribution by the second term. From this fact, and from the structure of the constants A and B defined at (3.10), it is clear that larger values of g tend to increase variance but decrease bias. (However, the manner in which this is achieved is perhaps a little crude relative to the smoother, non-dyadic local adaptability obtainable from a variable-bandwidth kernel density estimator. There, g is effectively arbitrary, not constrained to take integer values.)

In the case of the second spatially adaptive threshold rule, given at (3.11), each of the quantities p_0, g and C is a smoothing parameter. Particularly if C depends on location, as suggested in the paragraph following (3.11), this threshold rule grants greater flexibility in smoothing than that considered in the previous paragraph.

5. Piecewise-Smooth Densities

In the foregoing discussion it has been convenient to consider f to be an r-times differentiable function, without any discontinuities in $f^{(j)}$ for $0 \le j \le r-1$. However, all our results and conclusions about wavelet estimators carry over to the case of piecewise-smooth functions, where the class $\mathcal{F}_r(B)$ defined in Section 2.1 may be enlarged to

$$\begin{aligned}\mathcal{G}_r(B) = \Big\{ & f \ge 0 : \int f = 1;\ \text{for } 0 \le j \le r,\ f^{(j)} \text{ is well-defined and} \\ & \text{continuous at all but at most } B \text{ points, all lying within} \\ & (-B, B), \text{ at which both left and right derivatives exist;} \\ & \sup_x |f^{(j)}(x+)| \le B, \sup_x |f^{(j)}(x-)| \le B, \int |f^{(j)}| \le B \text{ for} \\ & 0 \le j \le r; f^{(r)} \text{ is monotone on } (-\infty, -B) \text{ and on } (B, \infty) \Big\}.\end{aligned}$$

The only change necessary is to ask that $2^q\, n^{-2r/(2r+1)} \to \infty$, which one may ensure by insisting that $q \sim C_2(\log n)(\log 2)^{-1}$ where $C_2 \in (2r(2r+1)^{-1}, 1)$ (instead of $C_2 \in (2r+1)^{-1}, 1)$, as was formerly the case. This alteration guarantees that the contribution from jump discontinuities to the series $\sum_{k \ge q} \sum_l b_{kl}^2$, which appears in the formula for mean integrated squared error (see (2.4)), is of smaller order than $n^{-2r/(2r+1)}$. Note that at discontinuities, b_{kl}^2 is of order p_k^{-1}, and that for each k the number of l's such that b_{kl} is affected by a discontinuity is uniformly bounded. A similar argument shows that the contribution of jump discontinuities to all other parts of the mean integrated squared error formula is of smaller order than $n^{-2r/(2r+1)}$.

The fact that our conclusions carry over to the case of piecewise differentiable f's indicates the considerable spatial adaptability of wavelet methods. By way of contrast, formula (3.6) for the mean integrated squared error of kernel estimators is not valid in the case of a discontinuous f, and in fact the optimal convergence rate in that context is an order of magnitude slower than for the wavelet estimators $\hat{f}_1$ and $\hat{f}_2$.

Acknowledgement

We are grateful to Professor I. M. Johnstone for helpful discussion. We also thank both referees for their comments which helped to improve exposition.

References

Donoho, D. L. (1992). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. Technical Report No. 403, Department of Statistics, Stanford University.

Donoho, D. L. and Johnstone, I. M. (1992a). Minimax risk over l_p-balls for l_q-error. Technical Report No. 401, Department of Statistics, Stanford University.

Donoho, D. L. and Johnstone, I. M. (1992b). Minimax estimation via wavelet shrinkage. Technical Report No. 402, Department of Statistics, Stanford University.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika **81,** 425–455.

Fan, J., Hall, P., Patil, P. and Martin, M. (1993). Adaptation to high spatial inhomogeneity based on wavelets and local linear smoothing. Research Report SMS–59–93, Centre for Mathematics and its Applications, The Australian National University.

Hall, P. and Patil, P. (1995a). On wavelet methods for estimating smooth functions. Bernoulli **1**, 41–58.

Hall, P. and Patil, P. (1995b). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. Ann. Statist. **23**, 905–928.

Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. Statist. Probab. Lett. **13**, 15–24.

Kerkyacharian, G. and Picard, D. (1993a). Density estimation by kernel and wavelet methods, optimality in Besov spaces. Manuscript.

Kerkyacharian, G. and Picard, D. (1993b). Linear wavelet methods and other periodic kernel methods. Manuscript.

Kerkyacharian, G. and Picard, D. (1993c). Introduction aux Ondelettes et Estimation de Densité, 1: Introduction aux Ondelettes et a l'Analyse Multiresolution. Lecture Notes.

Rosenblatt, M. (1971). Curve estimates. Ann. Math. Statist. **42**, 1815-1842.

Centre for Mathematics and its Applications, Australian National University.

(Received May 1994; accepted April 1995)